Two-tailed significance tests for 2×2 contingency tables

Two-tailed Significance Tests for 2x2 Contingency Tables: What is the Alternative?

Robin John Prescott

Centre for Population Health Sciences, Usher Institute, University of Edinburgh, UK robin.prescott@ed.ac.uk

Keywords

2x2 Contingency Tables
Two-tailed tests
Alternative hypotheses
Fisher's exact test

Abstract

Two-tailed significance testing for 2x2 contingency tables has remained controversial. Within the medical literature different tests are used in different papers and that choice may decide whether findings are adjudged to be significant or non-significant; a state of affairs that is clearly undesirable. In this paper it is argued that a part of the controversy is due to a failure to recognise that there are two possible alternative hypotheses to the Null. It is further argued that while one alternative hypothesis can lead to tests with greater power, the other choice is more applicable in medical research. That leads to the recommendation that, within medical research, 2x2 tables should be tested using double the one-tailed exact probability from Fisher's exact test or, as an approximation, the chi-squared test with Yates' correction for continuity.

1 Introduction

The 2x2 contingency table is almost the simplest data structure that one can encounter and yet there is no harmony among statisticians as to which two-sided test of significance is most appropriate. There are, of course, situations where a one-sided test may be more appropriate, but the analysis in this situation generates less controversy and will not be considered further in this paper. For discussion of the arguments in favour of one-sided or two-sided tests of significance, the reader is referred to Senn[1]. In relation to the two-tailed test, Martin Bland[2] has described acrimonious discussions which are still unresolved but are "generating almost as much heat as light". Although there are numerous methods that have been proposed as two-tailed significance tests for 2x2 tables, this paper will focus on four that are most commonly reported, as these will illustrate the point the author wishes to make. They are the Chi-squared test without continuity correction, the Chi-squared test with Yates' continuity correction, double the one-tailed exact probability from Fisher's exact test and the two-tailed exact probability from Fisher's exact test. To simplify the presentation, these will be also be referred to as Pearson's Chi-squared, Yates' Chi-squared, Fisher's Double P-value and Fisher's Added-tails.

In more extensive contingency tables, such as a general rxc table, statisticians are used to thinking about the most appropriate alternative hypothesis to the Null and choosing a test that is suitable for that alternative. If both rows and columns represent an ordered categorical variable but a general alternative that row and column variables are not independent is chosen, a chi-squared test with (r-1).(c-1) degrees of freedom will be used. However, if the alternative is that increasing values of the row variable are associated with increasing or decreasing values of the column variable then the method of choice will be the Jonckheere-Terpstra test[3,4].

It may not be immediately apparent but a similar choice is available for 2x2 contingency tables. The alternative hypothesis that is, de facto, in routine use, is that the row and column variables are not independent of each other. With that alternative, the author would have no hesitation is using the Pearson chi-squared test or the Fisher's Added-tails test, as these are undoubtedly more powerful than the other aforementioned tests. In medical research, however, there can be problems with the use of these tests and paradoxical findings can result, as noted in Section 2 and exemplified in Section 3.

2. Method

Let us suppose that our 2x2 table consists of membership of groups A and B for one variable and the outcomes of 'success' or 'failure' for the other variable. If the true probabilities of success in a wider population are P(A) and P(B) then we may define a composite alternative hypothesis that either $P(A) > P(B)$ or $P(B) > P(A)$. For overall statistical significance at the 5% level this leads us to perform two one-tailed tests at the 2.5% significance level. We note in passing that this corresponds to advice in the ICH E9 guideline[5], which states that "the approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings".

This alternative hypothesis to the Null immediately suggests the application of the exact one-tailed probability from Fisher's exact test. For a two-tailed test, the probability from the one-tailed test can

be doubled, as suggested by Fisher himself in a letter to David Finney in 1946[6]. Alternatively, the chi-squared test with continuity correction can be used, as this was derived by Yates[7] as an approximation to what later became known as Fisher's exact test.

Why should we prefer this alternative hypothesis to the Null over the conventional approach? As we will see in the examples in Section 3, the conventional approach can lead to rejection of the Null hypothesis while, simultaneously, it is not possible to conclude that either A is superior to B or vice versa. The inferences are therefore paradoxical. Such paradoxical inferences are excluded as a possibility for the composite alternative hypothesis, by its definition.

3. Examples

The first example is from a 9 year observational study to determine the efficacy of pre-hospital rapid sequence intubation (RSI) in paediatric traumatic brain injury[8]. One group of patients was transported to hospital by helicopter and received RSI, which was standard for a helicopter transfer. The controls were transported by ground transport and had no intubation performed. Outcome at 6 months was based on the Modified Glasgow Outcome Score and classified as favourable (mild disability or better) or unfavourable. The results in the subgroup of those with major trauma and for whom the status at 6 months was known are shown in Table 1. This example shows the most extreme disparity that can occur between the conventional alternative hypothesis and the composite alternative. The minimum expected value under the Null hypothesis is so small (2.4) that most researchers (but by no means all) will choose to report Fisher's exact test. Both the one-tailed exact probability and the two-tailed exact probability are identical at 0.031. One might anticipate that this could possibly cause unease in a researcher, primed to be cautious about the application of one-tailed tests. Nevertheless, even in this situation, reporting the two-tailed exact probability is common and has been seen by the author on numerous occasions when reviewing papers submitted to medical journals. At the conventional 5% level a statistically significant result will be reported if taking the traditional approach but, paradoxically, with a one-tailed probability of 0.031, superiority of one group over the other cannot be claimed. With the composite alternative we would report p=0.061 and conclude non-significance at the conventional 5% level.

The next example is taken from a randomised controlled trial to compare 2 kinds of breast implant following prophylactic mastectomy[9]. The incidence of severe complications leading to failed reconstructions were highlighted in the Abstract of the paper. That data is presented in Table 2. As in the first example, most researchers would choose to report Fisher's exact test, as the minimum expected value under the Null hypothesis is 4.3 and the total sample size is only 48. That yields a two-tailed exact probability of 0.068. However, the authors reported using a Pearson Chi-squared test. We see that is generates a result ($p = 0.047$) that is just statistically significant at the conventional 5% level. Disturbingly, the authors reported that the Chi-squared test gave $p < 0.0001$. However, even with accurate arithmetic, the authors would still have concluded that there was a statistically significant difference between the two implants with significantly more failures in the Protexa® group. That conclusion is in contrast to the one-tailed P-value of 0.052, which shows that the TiLoop® Bra is not superior to Protexa® with respect to failed breast reconstruction. Thus the conclusions would be paradoxical. With the composite alternative hypothesis, the Yates' Chi-squared test leads to a p-value close to that from Fisher's Double P-value of 0.103.

Both of the first two examples are based on relatively small sample sizes where we can expect the greatest differences between the use of the composite alternative hypothesis and the conventional alternative. The next example comes from a larger randomised controlled trial reported in the Lancet[10]. Patients discharged from a mental health crisis team all received a personal recovery

workbook, with randomisation into receiving this workbook by post or to ten sessions with a peer support worker who supported them in completing the workbook. The primary outcome of readmission to an acute service within a year is summarised in Table 3. With a minimum number of 64 in the 4 cells of the contingency table, few researchers would have reservations about using one or other of the versions of the Chi-squared test. Thus, using the conventional alternative hypothesis, the Pearson Chi-squared test would yield a statistically significant result at the 5% level. However, the conclusion that peer support produces a significantly better readmission rate could not be sustained as the one-tailed exact probability is 0.029. Therefore the conclusions would be paradoxical. It is also of interest to note the disparity between the p-values obtained using the Chi-squared test without the continuity correction and the Fisher's Added-tails test, despite the reasonably large sample sizes. It emphasises that the chi-squared test is not always a good approximation to Fisher's exact test, even when software sample size requirements are met. This contrasts with the close agreement between the Yates' Chi-squared test and Fisher's Double P-value.

The authors[10] did not analyse the data in Table 3 directly but based their inferences on a logistic model, adjusted for centre and clinical condition. They obtained an OR of 0.66 with 95% confidence limits from 0.43 to 0.99, and reported a p-value of 0.0438. For comparison, note that the OR, associated 95% confidence interval and p-value from Table 3, using the standard asymptotic formula is 0.67 (0.45, 0.99), p=0.046. The exact 95% confidence interval is (0.44, 1.01).

Do the different inferences, highlighted in the previous examples, make any difference in practice? The next example suggests that they might. Between 2005 and 2008 Ambulance Victoria undertook a randomised trial to compare urban road-based paramedic drug-assisted rapid sequence intubation (RSI) of patients with severe traumatic brain injury to transport and subsequent intubation in the hospital emergency department[11]. The primary outcome was the extended Glasgow Outcome Scale (GOSe)[12]. This is an 8 point scale with values ranging from dead (1) to normal (8). Analysis of the primary outcome variable by the Mann-Whitney U-test gave p=0.28. One of the secondary outcome was the proportion of patients with a good neurological outcome, defined as scores of 5-8 on the GOSe. The results are summarised in Table 4. The authors reported p = 0.046, using a Pearson Chi-squared test and the full Conclusion in the Abstract was " In adults with severe TBI, prehospital rapid sequence intubation by paramedics increases the rate of favorable neurologic outcome at 6 months compared with intubation in the hospital". The absence of statistical significance for the primary outcome and all other secondary outcomes were not mentioned in the Conclusion. Subsequently, the RSI protocol was implemented in Victoria for routine use by road-based paramedics in all patients with coma (Glasgow Coma Score ≤9) of both traumatic and non-traumatic causes[13].

In fact, the conclusion of superiority for RSI using this secondary outcome is not justified as the one-tailed exact probability is 0.03 and the inferences from the authors' analysis are paradoxical. Using the more appropriate composite alternative hypothesis, both the Yates' Chi-squared test and Fisher's Double P-value test yield p=0.06 and we would conclude that the association is non-significant at the 5% level. As the authors report that there were no significant differences in the primary outcome variable, in intensive care or hospital length of stay, or in survival to hospital discharge, one might speculate on the extent to which the method of analysis of one particular 2x2 table, from a secondary outcome variable, was responsible for influencing future policy. If a non-significant difference had been reported, in accordance with the suggestion in this paper, might the policy decision have been different?

4. Discussion

The use of statistical methods that can lead to paradoxical inferences is clearly sub-optimal in some sense. However, as shown above, if we apply two-tailed significance tests to a 2x2 contingency table without carefully considering the alternative hypothesis, this can happen. These paradoxes are avoided if a composite alternative hypothesis that one 'treatment' is superior to the other is applied. It leads, naturally, to the application of two one-tailed Fisher's exact tests at the 2.5% level to achieve an overall test at the 5% level of significance. This test is approximated very well by the chi-squared test with Yates' correction for continuity, but do we really need to use this approximation? Yates derived the test in 1934[7] when the computation of a one-tailed Fisher's exact test would be lengthy unless the sample sizes were very small. With computers now able to perform the computations almost instantaneously, surely, as a profession, we should be advocating routine use of Fisher's exact test in preference to approximations? This has been the approach of the author in producing guidelines for submissions to Gait and Posture[14].

Closely related to the choice of significance tests is the choice of methods for calculating confidence intervals for a summary measure. With 2x2 tables, there is a choice between the use of relative risks, odds ratios or the absolute difference in proportions. The standard methods that are available in software packages are invariably based on asymptotic results and are comparable to the use of an uncorrected chi-squared test to assess statistical significance. Thus the 95% confidence limits will exclude the Null value in the examples previously shown, creating another potentially paradoxical situation. This was illustrated with the conventional 95% confidence intervals for the odds ratio from Table 3, which excluded the Null value of 1. In contrast, the exact method, as expected, has a 95% confidence intervals that includes 1, and the conclusions are consistent with the test based on the composite alternative hypothesis. Therefore, the author contends that, in parallel with advocating the use of the Fisher's Double P-value test, we should also be advocating the use of exact or test-based confidence intervals for 2x2 contingency tables.

Particular care has to be taken in the calculation of confidence intervals for the absolute difference in proportions. Newcombe[15] has evaluated 11 methods and, for some, coverage probabilities can be poor, especially with small expected values in any cell of the 2x2 table. Newcombe found that the exact method generally performs as expected but notes the large amounts of computation time involved. He proposed a computationally simpler method based on the Wilson score method[16] and incorporating a continuity correction, with good coverage properties. This should be a sound alternative if the exact method is computationally infeasible. Although this method may not be available in all statistical software packages, it is available in some and it is an option within the TABLES statement of SAS® PROC FREQ through use of RISKDIFF (CL=NEWCOMBE CORRECT).

Despite the demonstrated paradoxes that can result from the conventional approach to significance testing in 2x2 tables, changing practice will not be easy. Most analyses of 2x2 contingency tables will not be made by researchers who read Statistics in Medicine. Many will be statistically naive and liable to follow the recommendations of the manuals of the statistical software that they use. It is difficult to see meaningful progress being made until the profession unites in the advice that it gives to the wider community that applies statistical techniques. It requires us to persuade the journals for which we review to modify their advice to authors. We particularly need to be pro-active in influencing the guidance that statistical software manufacturers give to their users, for the software undoubtedly influences the statistical methods that are used in the papers we referee.

The arguments advanced in this paper have been derived for 2x2 tables but they apply to all situations where there is a comparison between two groups. The central tenet is that in medical research, if we are applying a two-tailed test of significance, we should be seeking to differentiate

between three possible conclusions with our inferences, rather than a simple binary decision of acceptance or rejection of the Null hypothesis. If we reject the Null hypothesis we should either be able to claim that one group is superior to the other or vice versa. In most situations this will be guaranteed by the symmetry of the test as in, for example, the Student t-test. For any asymmetrical test, a composite alternative hypothesis based on superiority of one or other group should be used, effectively leading to the application of two one-tailed tests at the $\alpha/2$ level.

It must be recognised that the methods advocated in this paper come with a downside. The downside is that power is reduced compared to using a general alternative to the Null, and arguments about power and the size of the test have been central to many of the arguments about the best way to analyse 2x2 contingency tables. Despite this, the author believes that this is a necessary price, well worth paying, in order to obtain coherent inferences in medical research.

References

1. Senn SJ. Statistical Issues in Drug Development. Hoboken: Wiley; 2007.

2. Bland M. An Introduction to Medical Statistics. 4th ed. Oxford, Oxford University Press; 2015.

3. Terpstra TJ. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Indag Math. 1952; 14:327–333.

4. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. Biometrika 1954; 41:133–145.

5. ICH E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonised Tripartite Guideline. Stat Med. 1999; 18:1905–1942.

6. Yates F. Tests of significance for 2 x 2 contingency tables (with discussion) J R Stat Soc Ser A. 1984; 147:426-463.

7. Yates F. Contingency table involving small numbers and the $\chi 2$ test. J R Stat Soc Supp. 1934; 1:217–235.

8. Heschl S, Medley B, Andrew E, Butt W, Bernard S, Smith K. Efficacy of pre-hospital rapid sequence intubation in paediatric traumatic brain injury: A 9-year observational study. Injury 2018; 49: 916-920

9. Gschwantler-Kaulich D, Schrenk P, Bjelic-Radisic V, et al. Mesh versus acellular dermal matrix in immediate implant-based breast reconstruction - A prospective randomized trial. Eur J Surg Oncol. 2016; 42:665-671.

10. Johnson S, Lamb D, Marston L, et al. Peer-supported self-management for people discharged from a mental health crisis team: a randomised controlled trial. Lancet 2018; 392:409-418.

11. Bernard SA, Nguyen V, Cameron P, Masci K, et al. Prehospital rapid sequence intubation improves functional outcome for patients with severe traumatic brain injury: a randomized controlled trial. Ann Surg. 2010; 252(6):959-65.

12. Jennett B, Snoek J, Bond MR, Brooks N. Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. J Neurol Neurosurg Ps. 1981; 44(4):285–93.

13. Bernard SA, Smith K, Porter R, et al. Paramedic rapid sequence intubation in patients with non-traumatic coma. Emerg Med J. 2015; 32(1):60-4.

14. Prescott RJ. Editorial: Avoid being tripped up by statistics: Statistical guidance for a successful research paper. Gait Posture. doi: 10.1016/j.gaitpost.2018.06.172. [Epub ahead of print]

15. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998; 17:873-890.

16. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927: 22:209-212.

Table 1 Six month functional outcomes in children with major traumatic brain injury, initially treated with rapid sequence intubation (RSI) or no intubation[8]

	RSI	No intubation	Total
Favourable outcome	31	1	32
Unfavourable outcome	16	5	21
Total	47	6	53

Significance Test	Test Statistic	p
Pearson's Chi-squared	5.40	0.020
Fisher's Added-tails		0.031
Yates' Chi-squared	3.54	0.060
Fisher's Double P-value		0.061

Table 2 Complications resulting in implant loss in a randomised controlled trial to compare 2 types of breast implant following prophylactic mastectomy[9]

	Protexa®	TiLOOP® Bra	Total
Implant loss	7	2	9
No implant loss	16	23	39
Total	23	25	48

Significance Test	Test Statistic	p
Pearson's Chi-squared	3.96	0.047
Fisher's Added-tails		0.068
Yates' Chi-squared	2.62	0.105
Fisher's Double P-value		0.103

Table 3 Readmission to acute care over 1 year in patients discharged from a mental health crisis team and randomised to peer support or to a control group[10]

	Peer Support	Control	Total
Readmission	64	83	147
No readmission	154	133	287
Total	218	216	434

Significance Test	Test Statistic	p
Pearson's Chi-squared	3.98	0.046
Fisher's Added-tails		0.054
Yates' Chi-squared	3.59	0.058
Fisher's Double P-value		0.058

Table 4 Neurological outcomes at 6 months in patients with severe traumatic brain injury randomised to intubation by paramedics or intubation at hospital[11]

	Paramedic Intubation	Intubation at Hospital	Total
Favourable outcome	80	56	136
Unfavourable outcome	77	86	163
Total	157	142	299

Significance Test	Test Statistic	p
Pearson's Chi-squared	3.99	0.046
Fisher's Added-tails		0.049
Yates' Chi-squared	3.54	0.060
Fisher's Double P-value		0.060